AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Clinical Microbiology®

Streptococcus pseudopneumoniae: Use of Whole-Genome Sequences To Validate Species Identification Methods

Christian Salgård Jensen,[a] Katrine Højholt Iversen,[b] Rimtas Dargis,[a] Patricia Shewmaker,[c] Simon Rasmussen,[b] Jens Jørgen Christensen,[a,d] Xiaohui Chen Nielsen[a]

[a]The Regional Department of Clinical Microbiology, Slagelse, Region Zealand, Denmark

[b]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

[c]Centers for Disease Control and Prevention, Atlanta, Georgia, USA

[d]Institute of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark

ABSTRACT A correct identification of *Streptococcus pseudopneumoniae* is a prerequisite for investigating the clinical impact of the bacterium. The identification has traditionally relied on phenotypic methods. However, these phenotypic traits have been shown to be unreliable, with some *S. pseudopneumoniae* strains giving conflicting results. Therefore, sequence-based identification methods have increasingly been used for identification of *S. pseudopneumoniae*. In this study, we used 64 *S. pseudopneumoniae* strains, 59 *S. pneumoniae* strains, 22 *S. mitis* strains, 24 *S. oralis* strains, 6 *S. infantis* strains, and 1 *S. peroris* strain to test the capability of three single genes (*rpoB*, *gyrB*, and *recA*), two multilocus sequence analysis (MLSA) schemes, the single nucleotide polymorphism (SNP)-based phylogeny tool CSI phylogeny, a k-mer-based identification method (KmerFinder), average nucleotide identity (ANI) using fastANI, and core genome analysis to identify *S. pseudopneumoniae*. Core genome analysis and CSI phylogeny were able to cluster all strains into distinct clusters related to their respective species. It was not possible to identify all *S. pseudopneumoniae* strains correctly using only one of the single genes. The MLSA schemes were unable to identify some of the *S. pseudopneumoniae* strains, which could be misidentified. KmerFinder identified all *S. pseudopneumoniae* strains but misidentified one *S. mitis* strain as *S. pseudopneumoniae*, and fastANI differentiated between *S. pseudopneumoniae* and *S. pneumoniae* using an ANI cutoff of 96%.

KEYWORDS *Streptococcus*, mitis group, identification, methods, whole-genome sequencing, genotypic identification

The mitis group of the genus *Streptococcus* (MGS) consists of approximately 20 species (1). Most MGS strains are normal constituents of the human oral microbiome and rarely cause disease. One exception is *Streptococcus pneumoniae*, which is one of the most common causes of pneumoniae, meningitis, and otitis media (1).

The discrimination of *S. pneumoniae* from the remaining MGS strains has relied on phenotypic tests, with *S. pneumoniae* being optochin sensitive and bile soluble. In 2004, a new species, *Streptococcus pseudopneumoniae*, was described and characterized as being bile insoluble and sensitive to optochin when grown in ambient air but optochin resistant when grown in a CO_2-enriched atmosphere (2). Later studies have questioned the validity of these phenotypic traits for identification of *S. pseudopneumoniae* (3).

Sequencing of the 16S rRNA gene has been the standard reference identification method for most bacterial species. However, due to a high degree of interspecies similarity, 16S rRNA gene sequence analysis is not sufficient for the identification of MGS strains to the species level (4). Instead, multilocus sequence analysis (MLSA) schemes,

Citation Jensen CS, Iversen KH, Dargis R, Shewmaker P, Rasmussen S, Christensen JJ, Nielsen XC. 2021. *Streptococcus pseudopneumoniae*: use of whole-genome sequences to validate species identification methods. J Clin Microbiol 59:e02503-20. https://doi.org/10.1128/JCM.02503-20.

Editor Nathan A. Ledeboer, Medical College of Wisconsin

Address correspondence to Christian Salgård Jensen, csj@dadlnet.dk.

Received 2 October 2020
Returned for modification 15 October 2020
Accepted 9 November 2020
Accepted manuscript posted online 18 November 2020
Published 21 January 2021

consisting of several concatenated housekeeping genes, have been used. Two MLSA schemes have been reported for MGS, comprising six and seven housekeeping genes each (5–7). Due to the workload of doing PCR of and sequencing of 6 or 7 genes, these schemes are not feasible in clinical laboratories. Therefore, several single genes have been used for MGS species identification, including *recA* (8, 9), *rpoB* (10), and *gyrB* (11).

Next-generation sequencing (NGS) technology has rapidly expanded the sequences available (12). By using the long stretches of DNA obtained using NGS, it is possible to obtain a high degree of resolution between the strains tested (1, 13, 14). Furthermore, easy and accessible analysis methods for NGS data have emerged, making the phylogenetic analysis of data from unknown species faster (15–17).

The aim of this study was to elucidate the applicability of sequence-based identification methods within MGS, with the primary focus on identification of *S. pseudopneumoniae*. We tested single gene phylogeny, MLSA and multilocus sequence typing (MLST) schemes, average nucleotide identity (ANI), core genome analysis, and available online tools, with the use of CSI phylogeny as the inclusion identification method.

MATERIALS AND METHODS

Strains. Twenty *Streptococcus pseudopneumoniae* strains previously described by Arbique et al. (2) were obtained from the Centers for Disease Control and Prevention (CDC, Atlanta, GA, USA), and five additional clinical strains were tested and identified as suspected *S. pseudopneumoniae* strains with routine matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF MS) at the Regional Department of Clinical Microbiology, Slagelse, Region Zealand, Denmark.

For all strains, DNA was extracted using the Nextera DNA Flex microbial colony extraction protocol (Illumina, San Diego, USA). Library preparation for the 20 strains obtained from the CDC was with the Nextera XT DNA library preparation kit, the libraries were sequenced on an Illumina NextSeq 500 generating 150-bp paired-end reads and quality ensured using the Bifrost platform (https://github.com/ssi-dk/bifrost). Library preparation for the five clinical strains was with the Nextera DNA flex library preparation kit (Illumina); the libraries were sequenced on an Illumina MiSeq generating 150-bp paired-end reads and quality ensured using the CLC Genomics Workbench (Qiagen, Denmark). The raw data were assembled using SPAdes v. 3.13.0 (18), and assembly metrics were calculated using QUAST 5.0.2 (19).

Furthermore, we downloaded an additional 151 sequences (*S. pseudopneumoniae*, n = 39; *S. pneumoniae*, n = 59; *S. mitis*, n = 22; *S. oralis*, n = 24; *S. infantis*, n = 6; *S. peroris*, n = 1) from the NCBI database, see Data Set S1 in the supplemental material for details. Sequences were collected with an emphasis on having diversity regarding isolation sites and serotypes (20) and included the species type strains. To summarize, this study includes 176 genome sequences representing six species: *S. pseudopneumoniae* (n = 64), *S. pneumoniae* (n = 59), *S. mitis* (n = 22), *S. oralis* (n =24), *S. infantis* (n = 6), and *S. peroris* (n = 1).

CSI phylogeny. The single nucleotide polymorphism (SNP)-based CSI phylogeny Web application (https://cge.cbs.dtu.dk/services/CSIPhylogeny) was applied with default settings for species identification. In short, CSI phylogeny maps each contig against a chosen reference genome using BWA v. 0.7.2 (20). Afterwards, SNPs are called using SAMtools v. 0.1.18 (21), leaving out all SNPs in a 10-base vicinity of each other. CSI phylogeny returns a newick file for tree visualization. All strains were tested using their full assembled genomes, and sequences were uploaded in a fasta format (16). To avoid any bias toward a single species used in this study, we used *Streptococcus sanguinis* ATCC 10556^T as the reference genome.

Gene sequence extraction for single gene analysis and MLSA. From each genome, the following single genes were extracted for phylogenetic comparison: *recA*, *gyrB*, and *rpoB* for single gene phylogeny; *map*, *pfl*, *ppaC*, *pyk*, *rpoB*, *sodA*, and *tuf* for a seven-gene MLSA scheme (here, "MLSA scheme") (7); and *aroE*, *gdh*, *gki*, *recP*, *spi*, and *xpt* for a six-gene MLSA scheme (derived from the *S. pneumoniae* MLST scheme and therefore termed "MLST scheme") (5, 6).

To find and extract the genes from the genomes, we used an in-house Python script that uses previously published gene sequences as query sequences (see Data set S2). In short, the script makes a BLAST database for each strain, performs a dc-MEGABLAST search, and extracts the gene. The script returns a fasta file containing the hits present in the genome file tested. For the MLSA and MLST schemes, the Python script also returns the concatenated sequence for each strain.

All genes were aligned using MUSCLE v. 3.8.425 (22), and all gene alignments were visually inspected to ensure homology.

Core genome analysis. For the core genome analysis, genes were predicted and translated into amino acid sequences using prodigal v. 2.6.2 (21). The Bacterial Pan Genome Analysis tool (BPGA) version 1.3 (using USEARCH with a sequence identity cutoff of 0.5) was used to find the genes present in all strains (the core genome) used in this study (23). Genes were aligned using MUSCLE v. 3.8.425 (22). To ensure homology, all genes with <35% conserved sites were excluded from the analysis. All the remaining core genes were concatenated for further analysis.

Phylogenetic trees. A maximum likelihood tree based on the nucleotide sequences was constructed for each of the three single genes and the MLSA and MLST schemes using PhyML v 3.1, with SMS and bootstrapping (×100) (24).

For the core genome analysis, a tree based on the amino acid sequences was built using IQ-TREE, with the optimized maximum likelihood model, the gamma model, the -fast option applied, and bootstrapping (×100) (25).

All trees were visualized using the online tool iTOL (26).

KmerFinder. To perform KmerFinder, we used the Web application available at https://cge.cbs.dtu.dk/services/KmerFinder/ (software v. 3.0.2, accessed 29 October 2020) (15). The "standard output" was used, and the species of the "template" with the highest score was considered the result. If the template with the highest score was designated "streptococcus species," the result was categorized as "identification to genus level."

Average nucleotide identity. ANI was calculated using the tool fastANI v. 1.32 (17). All strain pairs were tested using the "many to many" method in fastANI and by using the "–matrix" option; results were obtained as a phylip-formatted lower triangular matrix. Earlier studies have shown that an ANI value of 94% to 96% corresponds to the recommended DNA-DNA hybridization species cutoff of 70%.

Data availability. The sequences of the 25 *S. pseudopneumoniae* strains sequenced during this study have been deposited in the NCBI database with BioSample numbers SAMN15921647 to SAMN15921671, under BioProject number PRJNA659631.

RESULTS

CSI phylogeny. CSI phylogeny was able to cluster all *S. pneumoniae*, *S. pseudopneumoniae*, and *S. oralis* strains in clearly distinct clusters together with a type strain, see Fig. 1. The *S. infantis* strains and the single *S. peroris* strain clustered together. The strains downloaded as *S. mitis* (except for one strain, see below), formed several subclusters. The subclustering of *S. mitis* has been described earlier (1). One strain, downloaded as *S. mitis* (*S. oralis* ATCC 6249) from the NCBI Web server, clustered in the *S. oralis* cluster. Further review of this strain confirmed that it was correctly identified as *S. oralis* by CSI phylogeny and that the identification had been corrected on the American Type Culture Collection (ATCC) website (https://www.lgcstandards-atcc.org/Products/All/6249.aspx) but not on the NCBI Web server (GenBank accession AEEN00000000.1). All other strains achieved the same species identification using CSI phylogeny as on the NCBI Web server.

Single gene analysis. We made phylogenetic trees based on the *recA*, *gyrB*, and *rpoB* gene sequences shown in Fig. 2 (see also Fig. S1 and S2 in the supplemental material). None of the single gene analyses were able to separate the strains into the phylogenetic groups defined by CSI phylogeny.

MLSA and MLST schemes. The phylogenetic tree based on the MLSA scheme showed clustering of *S. oralis*, *S. infantis*, and *S. pneumoniae* strains with their respective type strains (Fig. 3). The *S. peroris* type strain (ATCC 700780^T) clustered in the *S. infantis* cluster. Twenty of the 23 *S. mitis* strains formed one cluster around the type strain, and the remaining three *S. mitis* strains clustered in the *S. pseudopneumoniae* branch.

In contrast, the phylogenetic tree based on the MLST scheme showed that the *S. pseudopneumoniae* strains did not form a distinct cluster but had strains spreading to the *S. pneumoniae* and *S. mitis* branches (Fig. 4). *S. pneumoniae*, *S. mitis*, *S. oralis*, and *S. infantis* formed distinct clusters with their type strain using the MLST scheme.

Core genome analysis. Using BPGA and discarding all genes with <35% conserved sites, we found the core genome of the 176 strains to be 590 genes, with a total alignment length of 180,460 amino acids. As shown in Fig. 5, all *S. pneumoniae*, *S. pseudopneumoniae*, *S. mitis*, and *S. oralis* strains formed clusters together with their type strains; the *S. peroris* type strain lies inside the *S. infantis* cluster.

KmerFinder. We used the online platform of KmerFinder to identify our strains (Table 1) (16). All *S. pseudopneumoniae* and *S. pneumoniae* strains were correctly identified using KmerFinder. In contrast, one *S. mitis* strain (*S. mitis* SK1080) was misidentified as *S. pseudopneumoniae*, and one *S. mitis* strain and two *S. oralis* strains were only identified to the genus level. All *S. infantis* and *S. peroris* strains were only identified to the genus level.

ANI. The means and ranges of ANI values between different species can be found in Table 2, and all ANI values can be found in Data Set S3. Importantly, when testing the *S. pseudopneumoniae* strains, we did not see any overlap between the results

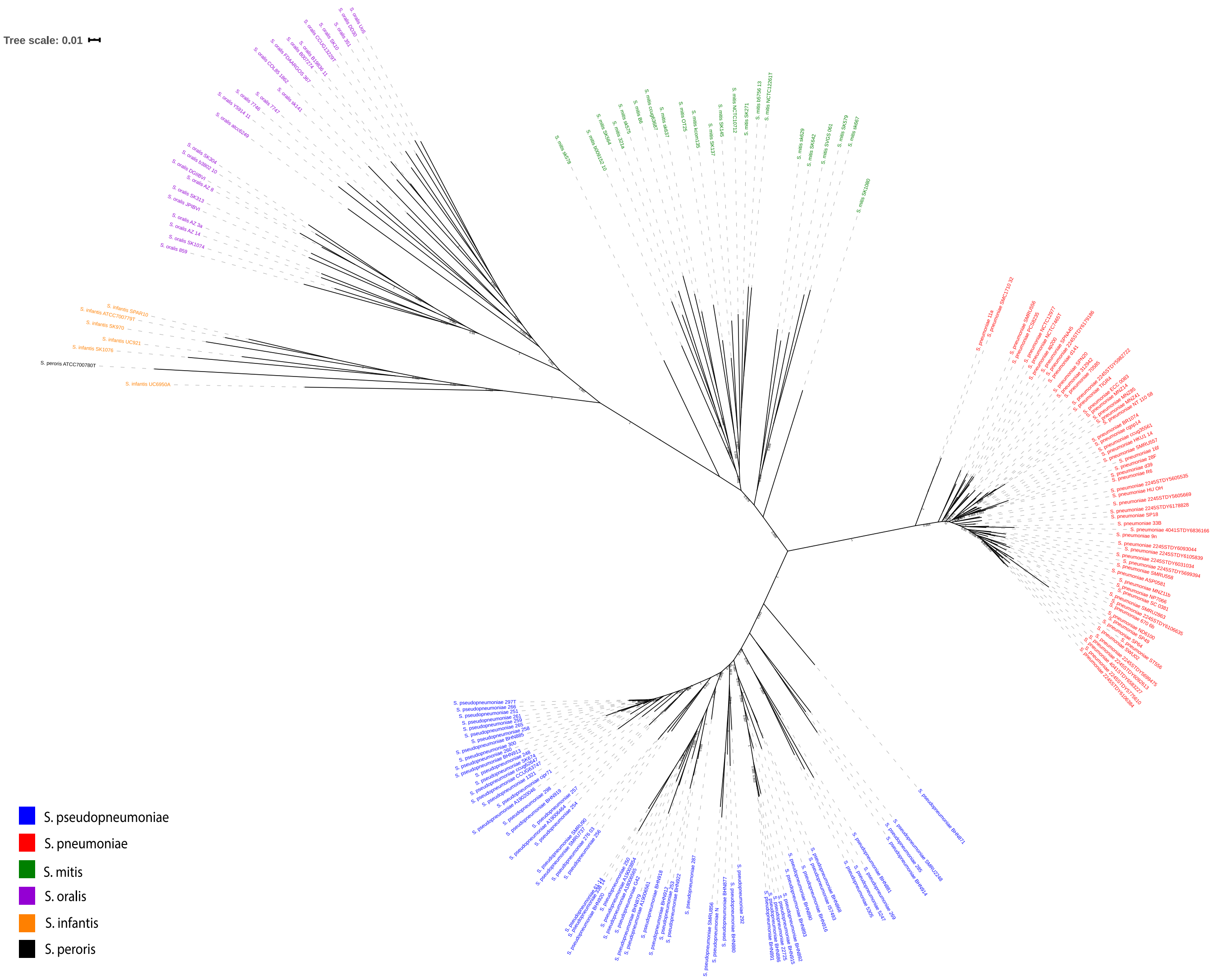

FIG 1 Phylogenetic tree made with 176 genomes belonging to the mitis group of the genus *Streptococcus* using the CSI phylogeny Web server (https://cge.cbs.dtu.dk/services/CSIPhylogeny). All strains were analyzed using whole-genome sequences and with the *S. sanguinis* type strain NCTC 7863 as the reference genome. Strains are colored according to the type strain contained in each cluster. The species designations were in concordance with the species annotation on the NCBI website, except for *S. oralis* ATCC 6249, which was annotated as a *S. mitis* (see text). Confidence scores of >0.9 are shown.

from *S. pseudopneumoniae* versus *S. pseudopneumoniae* and the results from *S. pseudopneumoniae* versus *S. pneumoniae*. Despite this nonoverlap, the distances between the lowest value when testing *S. pseudopneumoniae* versus *S. pneumoniae* and the highest value when testing *S. pseudopneumoniae* versus *S. pneumoniae* were 95.87% and 95.09%, respectively, which clearly reflects the close relationship between the two species. When testing *S. pseudopneumoniae* against *S. mitis*, *S. oralis*, *S. infantis*, and *S. peroris*, all had an ANI of <93.25%.

DISCUSSION

We used NGS sequences and bioinformatic tools to investigate identification schemes used to identify MGS to the species level, with an emphasis on *S. pseudopneumoniae*. Earlier studies have included up to 44 whole-genome sequenced *S. pseudopneumoniae* strains (3). In this study, we have included 64 whole-genome sequenced *S. pseudopneumoniae* strains, which, to our knowledge, is the largest collection to date. As an inclusion identification method, we used CSI phylogeny, which has previously been used for MGS with good results (23, 24). The sequence for one strain (*S. oralis* ATCC 6249) which was downloaded from the NCBI Web server as *S. mitis*, clustered in

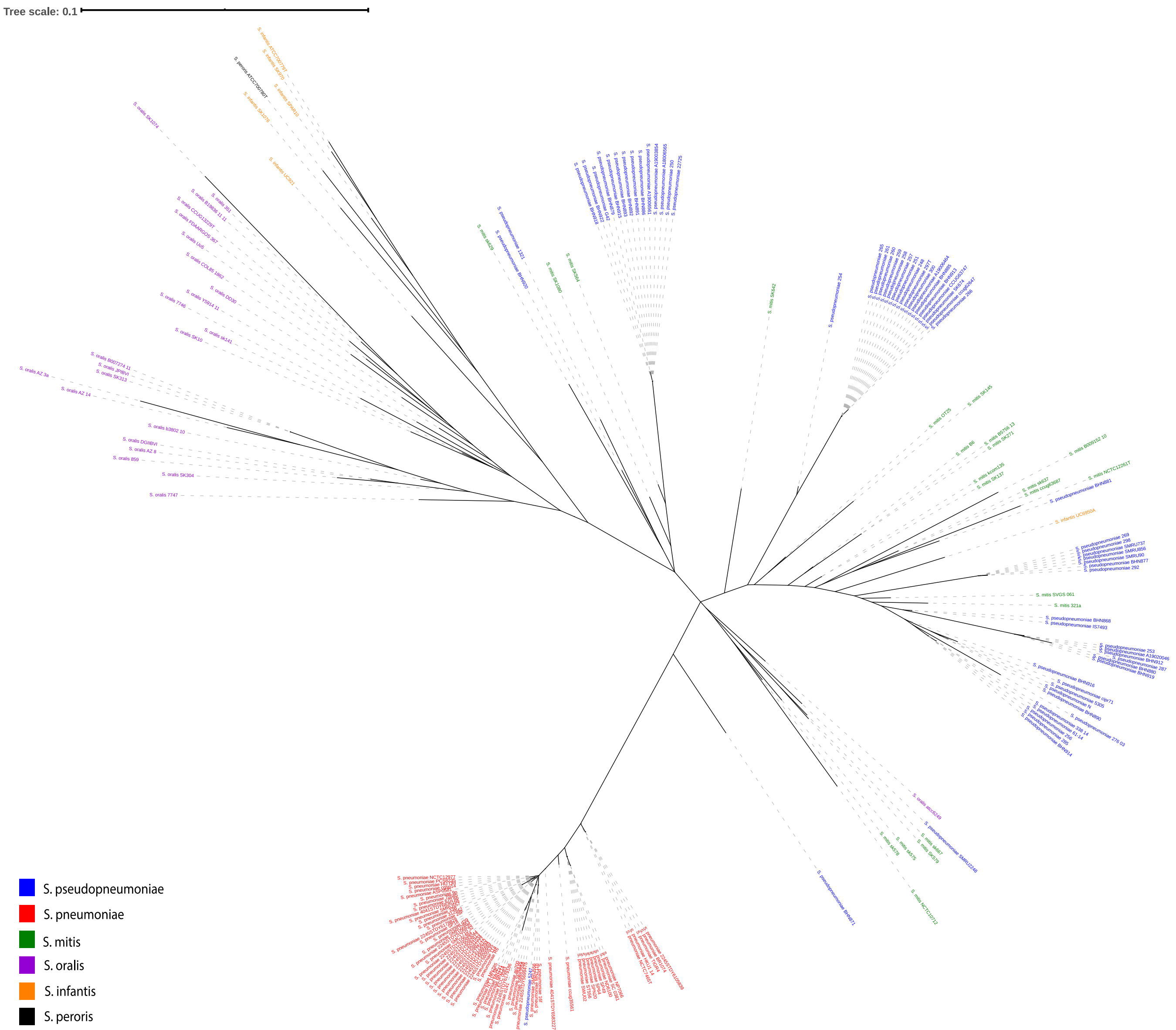

FIG 2 Maximum likelihood phylogenetic tree made using the *rpoB* gene. The gene sequences were extracted from whole-genome sequencing (WGS) and aligned using muscle; the tree was made using PhyML. Strains are colored according to the species obtained using the CSI phylogeny (Fig. 1).

the *S. oralis* cluster and should be considered *S. oralis*, which is in agreement with earlier studies (1) and the nomenclature on the ATCC website when we checked (https://www.lgcstandards-atcc.org/Products/All/6249.aspx).

Some studies have used single-gene sequence analyses to achieve a species identification of *S. pseudopneumoniae*. In a study including 11 *S. pseudopneumoniae* strains identified using phenotypic methods, *recA* was found to be able to discriminate *S. pneumoniae*, *S. pseudopneumoniae*, and *S. mitis* (9). Later, the *recA* gene was used to differentiate between *S. pseudopneumoniae* and the remaining MGS (8). A MALDI-TOF validation study used the *rpoB gene* to identify MGS, including 17 *S. pseudopneumoniae* strains. The use of the *rpoB* gene as the primary identification method was based on the gene having the best overall correlation between two MALDI-TOF systems and phenotypic methods (10). Zhou et al. used *gyrB* and 16S rRNA gene sequence analysis to identify 9 *S. pseudopneumoniae* strains and an additional 87 non-*S. pneumoniae* MGS strains and were able to cluster all MGS strains into distinct species clusters using *gyrB* (11). In our study, phylogenetic trees based on *recA*, *rpoB*, and *gyrB* gene

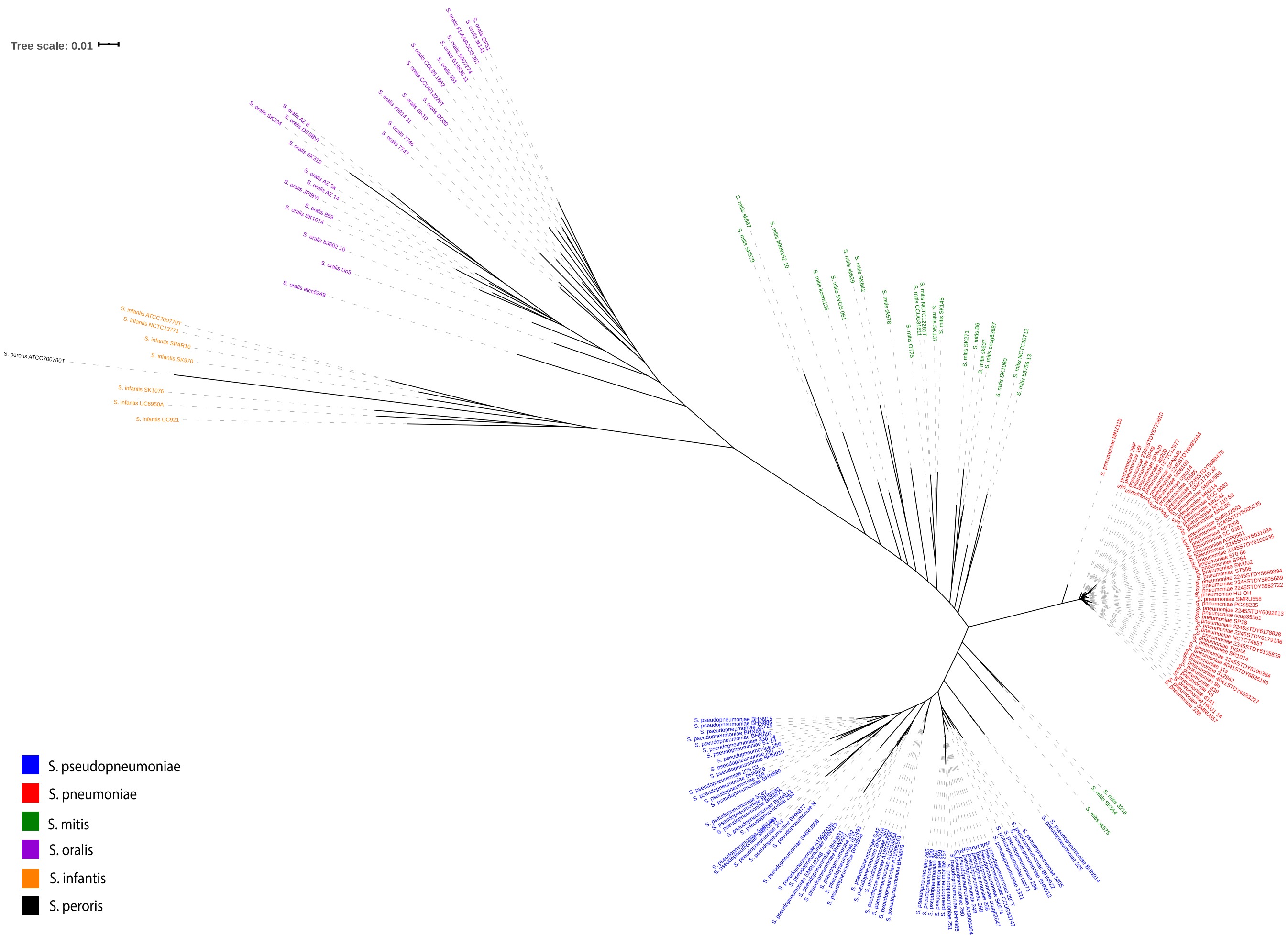

FIG 3 Maximum likelihood phylogenetic tree made using seven concatenated housekeeping genes included in the MLSA scheme (*map-pfl-ppaC-pyk-rpoB-tuf*) (7). The gene sequences were extracted from WGS and concatenated using an in-house Python script. A tree was made using PhyML with bootstrapping (×100). Strains are colored according to the species obtained using the CSI phylogeny (Fig. 1). Bootstrap values >90 are shown.

sequences revealed that none of the genes could be used to identify *S. pseudopneumoniae*, since no distinct clusters were formed.

Earlier studies have tested other genes for species identification within the MGS. Rasmussen et al. tested the genes *gdh*, *recA*, and *sodA* on a collection of 13 *S. mitis*, 26 *S. oralis*, 2 *S. infantis*, 18 *S. gordonii*, 20 *S. sanguinis*, and 1 *S. cristatus* strain. The study found that, especially, phylogeny inferred by the *gdh* gene often misidentified *S. oralis* strains (27). The *lytA* gene combined with restriction fragment length polymorphism analysis has been used for the identification of *S. pneumoniae* (28). Simões et al. found that some *S. pneumoniae* strains contained *lytA* genes that were different from the "typical" *S. pneumoniae lytA* gene and had the highest sequence similarity with the gene for the *S. pseudopneumoniae* type strain. Interestingly, Simões et al. also found two strains that clustered together with the *S. pseudopneumoniae* type strain using MLSA but contained a *lytA* gene with the highest sequence similarity to that for *S. pneumoniae* TIGR4 (29).

The results presented in this study and earlier studies emphasize that the identification of MGS should not be inferred from single genes, as it is often inaccurate, which is probably due to a high degree of interspecies horizontal gene transfer in this group (30).

To overcome the problem with horizontal gene transfer in MGS, an MLSA approach using concatenated sequences of six or seven housekeeping genes has been utilized. However, even the MLSA approach has been reported to be insufficient to identify some MGS strains. One study testing 132 MGS strains was unable to infer phylogeny

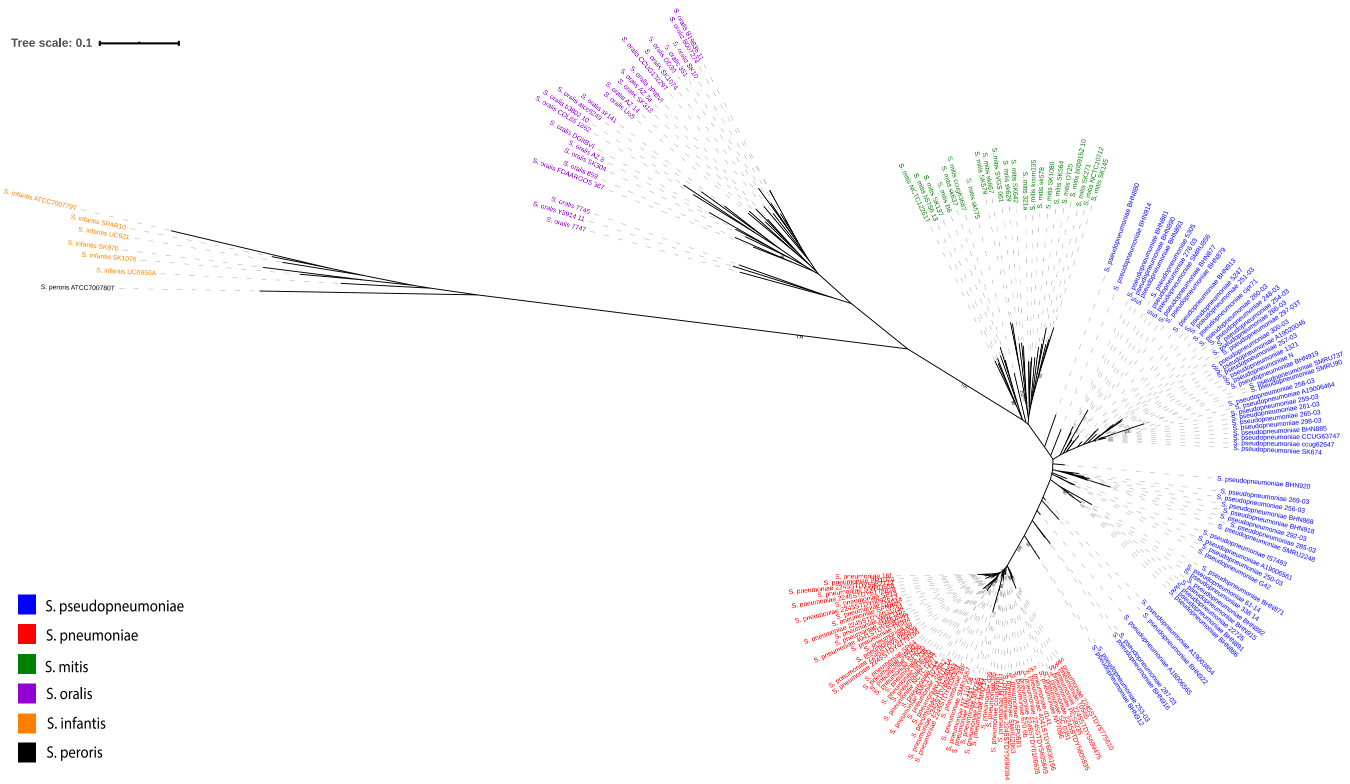
Tree scale: 0.1
S. oralis B19836 11
S. oralis B007274
S. oralis SK10
S. oralis 351
S. oralis DD30
S. oralis SK1074
S. oralis CCUG13229T
S. oralis JPIBVI
S. oralis AZ 3a
S. oralis SK313
S. oralis AZ 14
S. oralis Uo5
S. oralis sk141
S. oralis atcc6249
S. oralis b3802 10
S. oralis COL85 1862
S. oralis DGIIBVI
S. oralis AZ 8
S. oralis SK304
S. oralis 859
S. oralis FDAARGOS 367
S. oralis 7746
S. oralis Y5914 11
S. oralis 7747
S. infantis ATCC700779T
S. infantis SPAR10
S. infantis UC921
S. infantis SK970
S. infantis SK1076
S. infantis UC6950A
S. peroris ATCC700780T
S. mitis NCTC12261T
S. mitis b5756 13
S. mitis SK137
S. mitis B6
S. mitis sk637
S. mitis ccug63687
S. mitis sk575
S. mitis SK579
S. mitis sk667
S. mitis SVGS 061
S. mitis sk629
S. mitis SK642
S. mitis 321a
S. mitis kcom135
S. mitis sk578
S. mitis SK1080
S. mitis SK564
S. mitis OT25
S. mitis b009152 10
S. mitis SK271
S. mitis NCTC10712
S. mitis SK145
S. pseudopneumoniae BHN880
S. pseudopneumoniae BHN914
S. pseudopneumoniae BHN881
S. pseudopneumoniae BHN890
S. pseudopneumoniae BHN893
S. pseudopneumoniae 5305
S. pseudopneumoniae 276 03
S. pseudopneumoniae SMRU856
S. pseudopneumoniae BHN877
S. pseudopneumoniae BHN879
S. pseudopneumoniae BHN913
S. pseudopneumoniae 5247
S. pseudopneumoniae 251-03
S. pseudopneumoniae cipr71
S. pseudopneumoniae 260-03
S. pseudopneumoniae 248-03
S. pseudopneumoniae 254-03
S. pseudopneumoniae 266-03
S. pseudopneumoniae 297-03T
S. pseudopneumoniae 300-03
S. pseudopneumoniae A19020046
S. pseudopneumoniae 257-03
S. pseudopneumoniae 1321
S. pseudopneumoniae N
S. pseudopneumoniae BHN919
S. pseudopneumoniae SMRU737
S. pseudopneumoniae SMRU90
S. pseudopneumoniae 258-03
S. pseudopneumoniae A19006464
S. pseudopneumoniae 259-03
S. pseudopneumoniae 261-03
S. pseudopneumoniae 265-03
S. pseudopneumoniae 298-03
S. pseudopneumoniae BHN885
S. pseudopneumoniae CCUG63747
S. pseudopneumoniae ccug62647
S. pseudopneumoniae SK674
100
S. pseudopneumoniae BHN920
S. pseudopneumoniae 269-03
S. pseudopneumoniae 256-03
S. pseudopneumoniae BHN868
S. pseudopneumoniae BHN918
S. pseudopneumoniae 292-03
S. pseudopneumoniae 285-03
S. pseudopneumoniae SMRU2248
S. pseudopneumoniae IS7493
S. pseudopneumoniae A19006561
S. pseudopneumoniae 250-03
S. pseudopneumoniae G42
S. pseudopneumoniae BHN871
S. pseudopneumoniae 61-14
S. pseudopneumoniae 338 14
S. pseudopneumoniae BHN892
S. pseudopneumoniae BHN915
S. pseudopneumoniae 22725
S. pseudopneumoniae BHN891
S. pseudopneumoniae BHN886
S. pseudopneumoniae A19003854
S. pseudopneumoniae BHN922
S. pseudopneumoniae A18006565
S. pseudopneumoniae 287-03
S. pseudopneumoniae BHN916
S. pseudopneumoniae 253-03
S. pseudopneumoniae BHN912
95
S. pneumoniae 2245STDY5775610
🟦 S. pseudopneumoniae
🟥 S. pneumoniae
🟩 S. mitis
🟪 S. oralis
🟧 S. infantis
⬛ S. peroris

FIG 4 Maximum likelihood phylogenetic tree made using six concatenated housekeeping genes included in the MLST scheme (*aroE-gdh-gki-recP-spi-xpt*). The gene sequences were extracted from WGS and concatenated using an in-house Python script. A tree was made using PhyML with bootstrapping (×100). Strains are colored according to the species obtained using the CSI phylogeny (Fig. 1). Bootstrap values >90 are shown.

on two strains clustering at the root of the *S. pneumoniae* branch (31), and in another study, 28 of 634 strains did not cluster according to any recognized species (32). Our results are in concordance with these results, since some strains did not cluster together with their type strains. *S. pseudopneumoniae* can be especially difficult to identify using the MLST scheme, and some *S. mitis* strains could be misidentified as *S. pseudopneumoniae* using the MLSA scheme. Furthermore, when comparing with the phylogeny inferred by CSI phylogeny, some strains would probably be misidentified using the MLSA and MLST schemes.

Using core genome analysis, we were able to place all strains into the same defined clusters as when using CSI phylogeny. We used an amino acid-based approach for core genome analysis, which is very similar to the one used by Rasmussen et al. (27).

Two studies have used a nucleotide-based core genome approach. Garriss et al. (3) used a SNP-based method to infer phylogeny on 147 MGS strains, and Jensen et al. (1) used a method that relied on the alignment of 200-bp fragments of the *S. oralis* strain Uo5 against the 195 MGS strains tested. Despite the use of nucleotide sequences, different analysis methods, and different strain collections, the trees from the studies by Garriss et al. (3) and Jensen et al. (1) are remarkably similar to the tree obtained in our study. For example, all three core genome trees cluster the two *S. pseudopneumoniae* 338-14 and 61-14 strains and the two *S. mitis* SK575 and SK564 strains together, revealing the robustness of all three methods.

Interestingly, three *S. mitis* strains (321a, SK564, and SK575) were found to cluster outside the *S. mitis* cluster in the phylogenetic tree based on the MLSA scheme using the maximum likelihood method. These three strains have previously been found to cluster inside the *S. mitis* cluster using the minimum evolution method (1). The minimum evolution method is less computationally demanding than the maximum likelihood method but is reported to disregard much of the potential evolutionary information contained in sequences (33). To clarify whether the choice of phylogeny inference

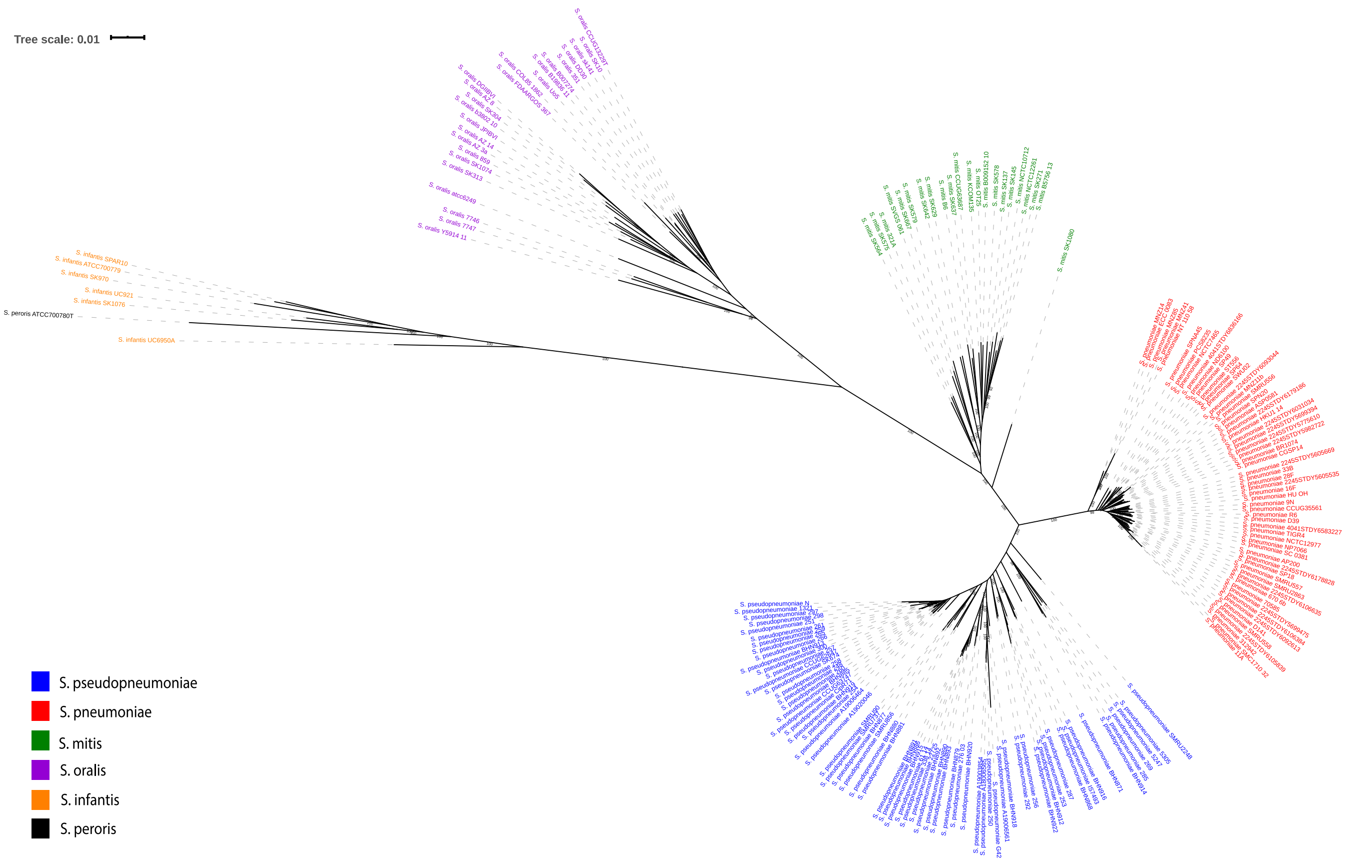

FIG 5 Maximum likelihood phylogenetic tree made with IQ-TREE using the concatenated sequences of 590 core genes, with a total amino acid alignment length of 180,460. Strains are colored according to the species obtained using the CSI phylogeny (Fig. 1). Bootstrap values >90 are shown.

method will influence the clustering pattern, we used MEGAX to construct minimum evolution trees with bootstrapping (×100) and compared them with those trees constructed using the maximum likelihood method (34) (see Fig. S3 to S8 in the supplemental material). We did not find any significant differences between the two methods using either single gene or core genome analysis.

KmerFinder compares k-mers of a query genome with k-mers from genomes contained in the database. It has been tested on a variety of publicly accessible genomes, including MGS, and was found to occasionally misidentify *S. mitis* as *S. pneumoniae* (15). The only misidentification experienced in the present study was *S. mitis* SK1080 being misidentified as a *S. pseudopneumoniae*. Interestingly, *S. mitis* SK1080 was an outlier in the *S. mitis* group using both core genome analysis and CSI phylogeny. Often, KmerFinder returns several matches with lower scores than the best match, and in our strain collection, three *S. mitis* strains have *S. pseudopneumoniae* as the second-

TABLE 1 KmerFinder results for 176 mitis group streptococci

	No. identified:		
Species (no. tested)	At the species level	At the genus level	Incorrectly[a]
S. pseudopneumoniae (64)	64	0	0
S. pneumoniae (59)	59	0	0
S. mitis (22)	20	1	1[b]
S. oralis (24)	22	2	0
S. infantis (6)	0	6	0
S. peroris (1)	0	1	0

[a] Species identification different from identification with CSI phylogeny.

[b] Identified as *S. pseudopneumoniae*.

TABLE 2 Pairwise ANI between the species in the study

	% ANI (mean [range])[a]				
Species	***S. pseudopneumoniae***	***S. pneumoniae***	***S. oralis***	***S. mitis***	***S. infantis***
S. pseudopneumoniae	97.54 (95.87–99.96)				
S. pneumoniae	94.48 (94.04–95.09)	98.45 (97.27–100.0)			
S. oralis	86.44 (85.97–86.94)	86.55 (86.01–87.2)	92.91 (90.42–98.82)		
S. mitis	92.12 (91.32–93.25)	91.73 (91.12–92.83)	86.84 (86.13–87.51)	93.31 (92.02–96.79)	
S. infantis	83.13 (82.39–83.92)	82.92 (82.3–83.7)	83.44 (82.94–84.09)	83.4 (82.76–84.3)	90.8 (88.16–94.92)
S. peroris[b]	82.64 (82.34–82.96)	82.4 (82.14–82.88)	82.58 (82.34–82.96)	82.72 (82.41–83.14)	88.34 (87.1–89.0)

[a] ANIs between all strains in one species against all strains in another species. The individual ANI values can be found in Data Set S3 in the supplemental material.
[b] Only one *S. peroris* strain was used in this study, and so an *S. peroris* versus *S. peroris* result is left out in the table.

best taxon match. Due to only one *S. pseudopneumoniae* strain being present in the database, most *S. pseudopneumoniae* strains have a second-best taxon match with *S. pneumoniae*. However, the score difference between best match and second-best taxon match was higher for *S. pseudopneumoniae* than for *S. mitis*, approximately 40,000 and 12,000, respectively. Interestingly, the three *S. mitis* strains (*S. mitis* SK564, *S. mitis* SK575, and *S. mitis* 321a) with *S. pseudopneumoniae* as second-best taxon match all cluster together in both the tree made by core genome analysis and the CSI phylogeny. As previously described, *S. mitis* contains several subclusters, and the misidentification and the lower distance to the second matched taxon probably reflects the lack of representatives within all *S. mitis* subclusters in the KmerFinder database.

ANI values between 94% and 96% have been shown to reflect the generally accepted DNA-DNA hybridization species boundary of 70% (1, 35). Previously, Jensen et al. used a BLAST-based method to calculate the ANI in a collection of 195 MGS strains, finding that only *S. pneumoniae*, *S. pseudopneumoniae*, and *S. gordonii* had a pairwise ANI of >96% (1). These results are comparable to the results found in the present study. However, we found two ANI values that were <96% when testing *S. pseudopneumoniae* versus *S. pseudopneumoniae*. However, this was a rare instance, since there were more than 2,000 ANI values calculated when testing *S. pseudopneumoniae* versus *S. pseudopneumoniae*. As shown before, most intraspecies ANI values of *S. oralis*, *S. mitis*, and *S. infantis* were <94%, making the 94% to 95% ANI value cutoff too high for species identification of these species. From our findings, an ANI cutoff of 96% would not misidentify an *S. pseudopneumoniae* strain as *S. pneumoniae* or as any other of the species tested in the present study.

In conclusion, we found that (i) core genome analysis clustered the *S. pseudopneumoniae* strains and the other examined MGS strains into distinct species clusters, similar to those found by CSI phylogeny, (ii) by using single gene analysis, misidentification of strains often occurs, so the method cannot be used for identification of *S. pseudopneumoniae*, (iii) MLSA and MLST schemes did not cluster all *S. pseudopneumoniae* strains into a single cluster, and misidentifications as *S. mitis* and *S. pneumoniae* may occur, (iv) KmerFinder was able to correctly identify all *S. pseudopneumoniae* strains, though one *S. mitis* strain was misidentified as *S. pseudopneumoniae*, and (v) fastANI can be used to discriminate *S. pseudopneumoniae* and *S. pneumoniae* and both from the remaining MGS using an ANI cutoff of approximately 96%.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.
SUPPLEMENTAL FILE 1, XLSX file, 0.2 MB.
SUPPLEMENTAL FILE 2, PDF file, 0.4 MB.

ACKNOWLEDGMENT

C.S.J. was funded by the Region Zealand Foundation for Health Research. K.H.I. and S.R. were supported by the Novo Nordisk Foundation grant NNF14CC0001.

None of the funders had any influence on the study or the conclusions.

REFERENCES

1. Jensen A, Scholz CFP, Kilian M. 2016. Re-evaluation of the taxonomy of the mitis group of the genus *Streptococcus* based on whole genome phylogenetic analyses, and proposed reclassification of *Streptococcus dentisani* as *Streptococcus oralis* subsp. *dentisani* comb. nov., *Streptococcus tigurinus* as *Streptococcus oralis* subsp. *tigurinus* comb. nov., and *Streptococcus oligofermentans* as a later synonym of *Streptococcus cristatus*. Int J Syst Evol Microbiol 66:4803–4820. https://doi.org/10.1099/ijsem.0.001433.
2. Arbique JC, Poyart C, Trieu-Cuot P, Quesne G, Carvalho M da GS, Steigerwalt AG, Morey RE, Jackson D, Davidson RJ, Facklam RR. 2004. Accuracy of phenotypic and genotypic testing for identification of *Streptococcus pneumoniae* and description of *Streptococcus pseudopneumoniae* sp. nov. J Clin Microbiol 42:4686–4696. https://doi.org/10.1128/JCM.42.10.4686-4696.2004.
3. Garriss G, Nannapaneni P, Simões AS, Browall S, Subramanian K, Sá-Leão R, Goossens H, de Lencastre H, Henriques-Normark B. 2019. Genomic characterization of the emerging pathogen *Streptococcus pseudopneumoniae*. mBio 10:e01286-19. https://doi.org/10.1128/mBio.01286-19.
4. Kawamura Y, Hou XG, Sultana F, Miura H, Ezaki T. 1995. Determination of 16S rRNA sequences of *Streptococcus mitis* and *Streptococcus gordonii* and phylogenetic relationships among members of the genus *Streptococcus*. Int J Syst Bacteriol 45:406–408. https://doi.org/10.1099/00207713-45-2-406.
5. Enright MC, Spratt BG. 1998. A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. Microbiology 144:3049–3060. https://doi.org/10.1099/00221287-144-11-3049.
6. Hanage WP, Kaijalainen T, Herva E, Saukkoriipi A, Syrjänen R, Spratt BG. 2005. Using multilocus sequence data to define the pneumococcus. J Bacteriol 187:6223–6230. https://doi.org/10.1128/JB.187.17.6223-6230.2005.
7. Bishop CJ, Aanensen DM, Jordan GE, Kilian M, Hanage WP, Spratt BG. 2009. Assigning strains to bacterial species via the internet. BMC Biol 7:3. https://doi.org/10.1186/1741-7007-7-3.
8. Harju I, Lange C, Kostrzewa M, Maier T, Rantakokko-Jalava K, Haanperä M. 2017. Improved differentiation of *Streptococcus pneumoniae* and other *S. mitis* group streptococci by MALDI biotyper using an improved MALDI biotyper database content and a novel result interpretation algorithm. J Clin Microbiol 55:914–922. https://doi.org/10.1128/JCM.01990-16.
9. Sistek V, Boissinot M, Boudreau DK, Huletsky A, Picard FJ, Bergeron MG. 2012. Development of a real-time PCR assay for the specific detection and identification of *Streptococcus pseudopneumoniae* using the *recA* gene. Clin Microbiol Infect 18:1089–1096. https://doi.org/10.1111/j.1469-0691.2011.03684.x.
10. Angeletti S, Dicuonzo G, Avola A, Crea F, Dedej E, Vailati F, Farina C, De Florio L. 2015. Viridans group streptococci clinical isolates: MALDI-TOF mass spectrometry versus gene sequence-based identification. PLoS One 10:e0120502. https://doi.org/10.1371/journal.pone.0120502.
11. Zhou M, Yang Q, Kudinha T, Zhang L, Xiao M, Kong F, Zhao Y, Xu Y-C. 2016. Using matrix-assisted laser desorption ionization-time of flight (MALDI-TOF) complemented with selected 16S rRNA and *gyrB* genes sequencing to practically identify clinical important *viridans* group streptococci (VGS). Front Microbiol 7:1328. https://doi.org/10.3389/fmicb.2016.01328.
12. Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333–351. https://doi.org/10.1038/nrg.2016.49.
13. Maiden MCJ, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND. 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol 11:728–736. https://doi.org/10.1038/nrmicro3093.
14. Pérez-Losada M, Cabezas P, Castro-Nallar E, Crandall KA. 2013. Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. Infect Genet Evol 16:38–53. https://doi.org/10.1016/j.meegid.2013.01.009.
15. Larsen MV, Cosentino S, Lukjancenko O, Saputra D, Rasmussen S, Hasman H, Sicheritz-Pontén T, Aarestrup FM, Ussery DW, Lund O. 2014. Benchmarking of methods for genomic taxonomy. J Clin Microbiol 52:1529–1539. https://doi.org/10.1128/JCM.02981-13.
16. Kaas RS, Leekitcharoenphon P, Aarestrup FM, Lund O. 2014. Solving the problem of comparing whole bacterial genomes across different sequencing platforms. PLoS One 9:e104984. https://doi.org/10.1371/journal.pone.0104984.
17. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114. https://doi.org/10.1038/s41467-018-07641-9.
18. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
19. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086.
20. Skov Sørensen UB, Yao K, Yang Y, Tettelin H, Kilian M. 2016. Capsular polysaccharide expression in commensal Streptococcus species: genetic and antigenic similarities to *Streptococcus pneumoniae*. mBio 7:e01844-16. https://doi.org/10.1128/mBio.01844-16.
21. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119.
22. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340.
23. Chaudhari NM, Gupta VK, Dutta C. 2016. BPGA- an ultra-fast pan-genome analysis pipeline. Sci Rep 6:24373. https://doi.org/10.1038/srep24373.
24. Lefort V, Longueville J-E, Gascuel O. 2017. SMS: Smart Model Selection in PhyML. Mol Biol Evol 34:2422–2424. https://doi.org/10.1093/molbev/msx149.
25. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274. https://doi.org/10.1093/molbev/msu300.
26. Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47:W256–W259. https://doi.org/10.1093/nar/gkz239.
27. Rasmussen LH, Dargis R, Højholt K, Christensen JJ, Skovgaard O, Justesen US, Rosenvinge FS, Moser C, Lukjancenko O, Rasmussen S, Nielsen XC. 2016. Whole genome sequencing as a tool for phylogenetic analysis of clinical strains of mitis group streptococci. Eur J Clin Microbiol Infect Dis 35:1615–1625. https://doi.org/10.1007/s10096-016-2700-2.
28. Llull D, López R, García E. 2006. Characteristic signatures of the *lytA* gene provide a basis for rapid and reliable diagnosis of *Streptococcus pneumoniae* infections. J Clin Microbiol 44:1250–1256. https://doi.org/10.1128/JCM.44.4.1250-1256.2006.
29. Simões AS, Tavares DA, Rolo D, Ardanuy C, Goossens H, Henriques-Normark B, Linares J, de Lencastre H, Sá-Leão R. 2016. *lytA*-based identification methods can misidentify *Streptococcus pneumoniae*. Diagn Microbiol Infect Dis 85:141–148. https://doi.org/10.1016/j.diagmicrobio.2016.03.018.
30. Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli SV, Oggioni M, Dunning Hotopp JC, Hu FZ, Riley DR, Covacci A, Mitchell TJ, Bentley SD, Kilian M, Ehrlich GD, Rappuoli R, Moxon ER, Masignani V. 2010. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol 11:R107. https://doi.org/10.1186/gb-2010-11-10-r107.
31. Rolo D, S Simões A, Domenech A, Fenoll A, Liñares J, de Lencastre H, Ardanuy C, Sá-Leão R. 2013. Disease isolates of *Streptococcus pseudopneumoniae* and non-typeable *S. pneumoniae* presumptively identified as atypical *S. pneumoniae* in Spain. PLoS One 8:e57047. https://doi.org/10.1371/journal.pone.0057047.
32. Scholz CFP, Poulsen K, Kilian M. 2012. Novel molecular method for

identification of *Streptococcus pneumoniae* applicable to clinical microbiology and 16S rRNA sequence-based microbiome studies. J Clin Microbiol 50:1968–1973. https://doi.org/10.1128/JCM.00365-12.

33. De Bruyn A, Martin DP, Lefeuvre P. 2014. Phylogenetic reconstruction methods: an overview. Methods Mol Biol 1115:257–277. https://doi.org/10.1007/978-1-62703-767-9_13.

34. Stecher G, Tamura K, Kumar S. 2020. Molecular Evolutionary Genetics Analysis (MEGA) for macOS. Mol Biol Evol 37:1237–1239. https://doi.org/10.1093/molbev/msz312.

35. Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A 106:19126–-19131. https://doi.org/10.1073/pnas.0906412106.